Mass Spectra Deconvolution of Gaseous Mixtures Containing Volatile Organic Compounds

D. Nikolić[1] and S. M. Madzunkov[2], M. Darrach[3]
California Institute of Technology, Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, CA, 91109

The Spacecraft Atmosphere Monitor (S.A.M.) analyzes all gaseous pollutants in the 1-150Th mass-to-charge range based on their positive-ion mass spectra. Of particular interest is the Trace Gas Analysis (TGA) mode of operation in which S.A.M detects minute amounts of volatile organic compounds (VOC). Pollutant component of the ambient air is sampled on demand in twenty full mass spectra per second. All spectra is accumulated in two-second intervals by an on-board Field Programmable Gate Arrays (FPGA) unit. Mass spectra represents the number of detected fragment ions in a given mass-to-charge ratio channel. Electron impact ionization of neutral molecules results in creation of numerous fragment ions and most of organic compounds will contribute several identical fragments. Despite the presence of these molecular isobars, we developed the deconvolution algorithm capable of identifying target species based on their characteristic fragmentation patterns. We investigate the efficiency of deconvolution algorithm as a function of mass resolution with which mass spectrum is acquired. Higher the mass resolution, better the chances are to distinguish between stereoisomers using their fragmentation propensities, but also results in higher data volumes to be processed by a set of small footprint software stacks hosted by an on-board computer. Finding the balance between deconvolution accuracy and generated data volume under time constrains and limited computing resources is the main topic of this study.

Nomenclature

GC	=	Gas Chromatograph	S.A.M.	=	Spacecraft Atmosphere Monitor
ISS	=	International Space Station	QIT	=	JPL Quadruple Ion Trap
JPL	=	Jet Propulsion Laboratory	*rf*	=	Radio Frequency
MCA	=	Major Constituents Analysis	SMAC	=	Spacecraft Max. Allowable
MEMS	=	Micro-Electro-Mechanical System	Conc.		
MS	=	Mass Spectrometer	TGA	=	Trace Gas Analysis
EII	=	Electron Impact Ionization	VOC	=	Volatile Organic Compound
CITA	=	Computational Ion Trap Analyzer	*m*/*q*	=	mass-to-charge quotient
PC	=	Preconcentrator	Th	=	Thomson

[1] Technologist, Group 389T, M/S 306-392.
[2] Senior Technologist, Group 389T, M/S 306-392.
[3] Senior Technologist, Group 389T, M/S 306-392.

I.Introduction

THIS progress report summarizes the initial development of Major Constituents Analyzer (MCA), software for the automated interpretation of mass spectrometer (MS) data delivered by the S.A.M.'s QITMS sensor module, which is currently under a development as a technological demonstration for the Internationl Space Station (ISS). The S.A.M. consists of a gas chromatograph (GC) unit integrated with a microfabricated preconcentrator (PC) and interfaced with the QITMS sensor. For further details on engineering/electronics architecture and baseline performance requirements please refer to the latest S.A.M. progress report. We repeat here only a novel aspect related to the Red Pitaya, an affordable open source development board managed by Debian GNU/Linux distribution. Powered currently by Zynq 7010 SoC, which combines a FPGA platform and a Cortex A9 dual core processor, the Red Pitaya enables a relatively inexpensive capability of onboard inflight analysis. Here we discuss only the efficiency of the algorithm used for the determination of the relative abundances of the 22 compounds listed in Table , to be tested for while the S.A.M operates in the TGA mode. In this mode of operation cabin air is circulated through PC for 10 minutes where any VOCs are adsorbed. The VOCs are then thermally desorbed and flushed into the QITMS sensor using H_2 carrier gas via the GC micro-column. Any potential VOCs are ionized, fragmented, and confined inside the QITMS and then detected with respect to the mass-to-charge ratio (m/q). Electron impact ionization (EII) fragmentation patterns for these compounds were adopted from NIST and used to create initial ion cloud compositions for various VOC mixtures. These initial ion clouds are then propagated in time using the Computational Ion Trap Analyzer (CITA) program with the quadrupole-only (end caps grounded) 1MHz *rf* scan function capable of trapping ions with $1 < m/q < 361$ Th with unit mass resolution. The trapping efficiency of the QITMS under a wide range of operating conditions will be reported elsewhere. Here we are primarily concerned with how initial uncertainties due to counting statistics alone propagate and accumulate in the final values for the relative abundances of all the VOCs of interest.

For example, Fig. illustrates mass spectra of all the compounds from Table when they simultaneously enter the analyzed gas mixture in equal parts. It is evident that at unit mass-to-charge resolution, a given mass channel may contain contributions of several different compounds, each providing its own fragment of a given mass with substantially different probability as determined by the EII cross sections at 75eV. Namely, the three isomeric forms of Xylene can be distinguished by the designations ortho- (o-), meta- (m-), and para- (p-), depending on which carbon atoms in the benzene ring bind the two methyl groups. Even though these three isomers have the same chemical formula, their EII fragmentation patterns differ to a certain extent. Despite the fragment C_7H_7 being the most abundant in all three Xylene isomers, the fragment C_3H_3 in m-Xylene is more than twice abundant than in the other two isomers. We would like to explore the notion that this particular difference in the EII fragmentation pattern of Xylene isomers can aid us in setting them apart when analyzing the unit mass resolution mass spectra. Another example of the complexity of the task at hand is the case of Cyclopentasiloxane (decamethyl compound-11), which can contribute with low probability a staggering 43 fragments at mass 207 Da as opposed to Cyclotrisiloxane (hexamethyl, compound-22) which contributes almost exclusively to the same mass channel with only two fragments. In order to identify all compounds of interest in the sample gas mixture, we use the multi-dimensional Monte-Carlo random walk simulation algorithm to convert mass-spectral line intensities (Fig.) into the relative abundances. The mixing ratios of mass spectra shown in Fig. can range from the case of minimal entropy with a single compound, to the case of maximum entropy with equipartial contribution of all compounds. The former case corresponds to ideal operation of the MEMS PC module when the elution chromatogram contains all the VOCs of interest clearly separated in time. The latter case would resemble a failure mode where the MEMS PC absorbs all the VOCs and releases them at the same into the QITMS where they fail to separate in time. We would like to use this latter test case for benchmarking the maximum entropy mode of the random walk algorithm.

II.Results and Discussion

Table 1 Error! No sequence specified.**: List of compounds used in simulated TG analysis.**

(c)	$A_{(c)}$	name	formula	(c)	$A_{(c)}$	name	formula
1	0.36	Silanol, trimethyl-	$C_3H_{10}OSi$	**12**	1.05	Propylene Glycol	$C_3H_8O_2$
2	0.32	2-Propenal-	C_3H_4O	**13**	0.44	Hexane	C_6H_{14}
3	0.06	Cyclotetrasiloxane, octamethyl-	$C_8H_{24}O_4Si_4$	**14**	0.02	Ammonia	H_3N
4	0.45	Formaldehyde	CH_2O	**15**	0.54	Methyl Alcohol	CH_4O
5	0.17	Perfluoropropane	C_3F_8	**16**	1.00	Acetaldehyde, tetramer	$C_8H_{16}O_4$
6	0.05	Methylene Chloride	CH_2Cl_2	**17**	0.51	Benzene, 1,4-dimethyl-(p-Xylene)	C_8H_{10}
7	0.49	Benzene, 1,3-dimethyl-(m-Xylene)	C_8H_{10}	**18**	0.52	Benzene, 1,2-dimethyl-(o-Xylene)	C_8H_{10}
8	0.45	Toluene	C_7H_8	**19**	0.12	Benzene	C_6H_6
9	0.37	Acetone	C_3H_6O	**20**	0.54	1-Butanol	$C_4H_{10}O$
10	1.02	Isopropyl Alcohol	C_3H_8O	**21**	0.87	Ethyl alcohol	C_2H_6O
11	0.04	Cyclopentasiloxane, decamethyl	$C_{10}H_{30}O_5Si_5$	**22**	0.07	Cyclotrisiloxane, hexamethyl-	$C_6H_{18}O_3Si_3$

To generate the initial ion cloud that will be confined inside the QITMS, we use the *TrapParticle* tool from the CITA suite of codes. We first generate N_{ion} = 1 million ions that are randomly sampled from EII distributions of all possible fragments originating from each chemical compound. In the case of equipartial mixture, these ions will populate the most prominent mass-to-charge (m/q) peaks found in Fig. , and less probable fragments will be suppressed due to insufficient count statistics. This particular test case addresses the issue of maximum similarity among various compounds, which is also enhanced by the restriction set forth by unit mass resolution - fine differences in masses of neighboring isobars will disappear when these mass peaks blend into 1 Da wide mass bins. At this low mass resolution the only way to enhance the dissimilarity between compounds is to increase the counting statistics so that less probable EII fragments appear in the mass spectrum and set apart chemically similar compounds. Therefore, we repeat the above analysis for cloud sizes of N_{ion} = 10 and N_{ion} = 100 million ions.

A. Overview of Computational Algorithm

Governed by fragment distributions $\acute{f}^{(c)} = \sum_m \alpha_m^{(c)} \acute{e}_m$ for select compounds (c), each spanning the mass range of $1 \le m \le 361$Da, and each with the unit baseline mass resolution of $\Delta m = 1$Da, every mass channel m is assigned its eigenvector $\acute{e}_m$ such that $0 \le \alpha_m^{(c)} \le 1$ determines the probability ($\sum_m \alpha_m^{(c)} = 1$) with which a compound (c) will contribute to this mass channel. The EII fragmentation probabilities, $\alpha_m^{(c)}$, are known in advance

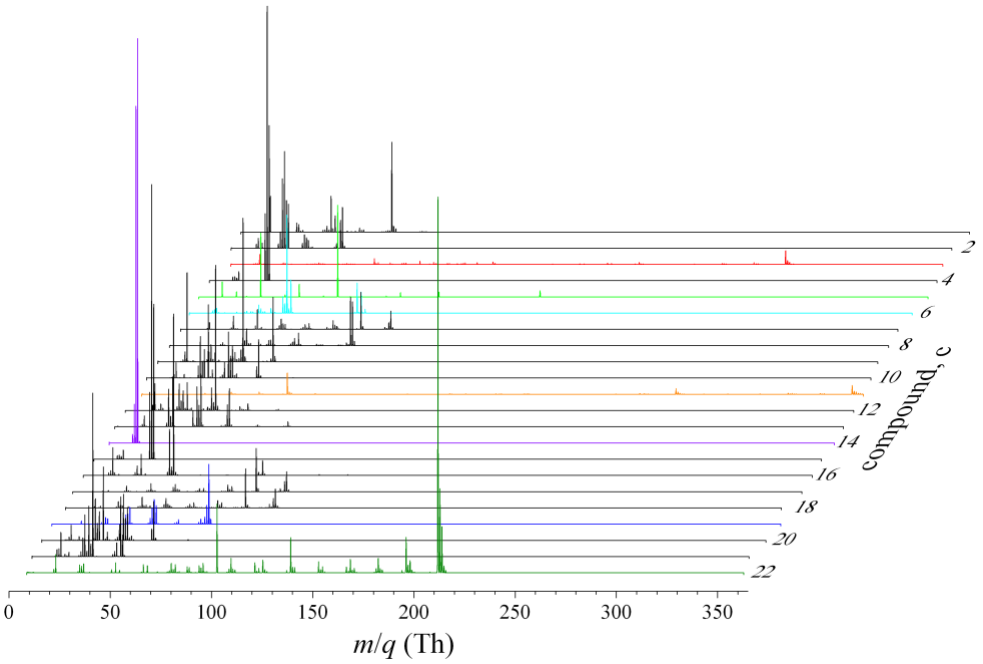

Figure 1 Error! No sequence specified.**: Simulated mass spectrum of VOCs found in Table 1.** *Fragmentally least similar compounds are given in color.*

for every compound of interest and for a given baseline mass resolution Δm only a few are non-zero. Furthermore, for each compound (c), we eliminate the cross-talk between mass channels m and m' by assuming the orthogonality of the compound's fragment space, that is $\acute{e}_m \cdot \acute{e}_{m'}$ is equivalent to the identity matrix. This assumption holds for unit mass resolution of QIT operating at low pressures (1E-7 Torr), where ion collisions with the neutrals are rare events. However, for sub-unit resolutions and higher pressures (1E-5 Torr), the $\acute{e}_m \cdot \acute{e}_{m'}$ must be a band matrix to properly account for the cross-talk between neighboring mass channels, $m' = m \pm \Delta m$. As a consequence of $\acute{e}_m \cdot \acute{e}_{m'}$ being equivalent to the identity matrix, the fragmental similarity matrix $A_{c,c'} = \sum_m \alpha_m^{(c)} \alpha_m^{(c')}$ is a sparse matrix. Its column-wise sums, $A^{(c)} = \sum_{c'} A_{c,c'}$, where $c' \ne c$, are reported in Table and are indicators of the uniqueness of the mass spectrum of each compound. Compounds with the smallest $A^{(c)}$ values (c = 14, 11, 6, 3, 22, 19, 5, given in color in Fig.) can form a computationally efficient subset of chemical compounds to initially span the entire experimental mass range.

For a known gas mixture we can generate a synthetic "experimental" mass spectrum $\acute{R}^{exp}$as a weighted sum of all mass spectra found in Fig. , such that the $N_m = \acute{R}^{exp} \cdot \acute{e}_m$ is the number of counts in a given mass channel m and

$N_{ion} = \sum_m N_m$ is the total number of counts (detected ions) in the experimental mass spectrum. In order to recover this predefined set of weighing coefficients, $\left\{\eta_{exp}^{(c)}\right\}_{c=1}^{22}$, the experimental mass spectrum $\acute{R}^{exp}$ is modeled by a trial distribution $\acute{R} = \sum_c \eta^{(c)} \acute{f}^{(c)}$ in which mixing coefficients, $0 \le \eta^{(c)} \le N_{ion}$, represent an absolute abundance of a target compound (c). For every mass channel m we can optionally assign the uncertainty in the counting statistics, $\sigma_m = \sqrt{N_m}$, or $\sigma_m = 1$ if no counts are detected. In solving the similarity problem between the two composite mass spectra, $\acute{R}$ and $\acute{R}^{exp}$, we minimize the merit function $\Delta R = min\left\|\acute{R} - \acute{R}^{exp}\right\|_2$ with respect to $\eta^{(c)}$ using the constrained least-square random walk method. Examining the explicit form for the square of the residual function, $\Delta R^2 = \sum_{c,c\prime} A_{c,c\prime}\left(\eta^{(c)} - \eta_{exp}^{(c)}\right)\left(\eta^{(c\prime)} - \eta_{exp}^{(c\prime)}\right)$, one can note that fragmentally similar compounds (large $A_{c,c\prime}$ values) contribute the most. We recommend that the minimization algorithm starts with a subset of compounds having smallest $A^{(c)}$ values, then gradually expand this subset to include the remaining compounds that need to be tested as well. The iterative random-walk procedure starts with equal weights $\eta^{(c)}$ for each compound (c) and successively updates their values $\eta^{(c)} \to \eta^{(c)} \pm \delta$ with fixed step size δ such that the distance ΔR corresponds to the global minimum.

In the event that the convergence rate for minimizing ΔR is slow or unstable due to the presence of multiple local minima, we introduce the exponential relaxation probability $P(\Delta R) = exp\,(-\lambda \Delta R)$ such that an arbitrary scale length λcontrols the rate of relaxation. Eventhough the new set of weight coefficients $\left\{\eta^{(c)}\right\}_{c=1}^{22}$ may not reduce the current residual ΔR value, we may still adopt it, provided a random draw does not exceed the probability $P(\Delta R)$. In this way any stagnation in the convergence rate is more likely to be disrupted for smaller ΔR values when the system is near the global minimum. Close to local minima, where the residual ΔR has presumably larger values, any repeated increase in ΔR outside the prescribed tolerance is likely to be rejected leading to either the time out (no solution found) or a restart of the algorithm using a new random selection of compounds (c). If convergence is reached to within the prescribed tolerance, a report on the relative abundances of all the investigated compounds is generated.

Overall success of the random walk algorithm depends on the completeness of the training set of compounds, and is often a trade-off between the execution time and desired sensitivity to the trace amount of VOCs. One way of achieving the optimal size of the training set is to randomly select least similar compounds, one by one, from the library of all possible chemicals present onboard the ISS. For each such compound, its maxiumum $\eta_{max}^{(c)}$ abundance supported by the experimental mass spectrum $\acute{R}^{exp}$ is found by minimizing the single-compound residuals $\Delta R^{(c)} = min\left\|\eta_{max}^{(c)} \acute{f}^{(c)} - \acute{R}^{exp}\right\|_2$. These residuals are then ranked and only the few lowest are retained as a starting subset of most probable and least similar compounds. Their mixing coefficients are now simultaneously adjusted for possible overlaps using the relaxed set of constraints, $0 \le \eta^{(c)} \le \eta_{max}^{(c)}$, by minimizing this time around the composite merit function ΔR. Any additional compounds that are about to be added into the system must be justified by a measurable reduction of the ΔR value. In this way, the number of random combinations of all compounds to be tested is significantly smaller and faster execution times are possible.

Our main concern in this phase of development is the reliability of the random-walk procedure under various counting statistics uncertainties. This translates to the shortest sampling times of pre-concetrated cabin atmosphere that will still yield a sufficient accuracy in the relative abundances of VOCs. Based on annual pumping requirements and the capacity of the ion/getter pump, the continuous MCA mode of operation will report relative abundances of cabin air components (CH_4, H_2O, N_2, O_2, Ar, and CO_2) every 2 seconds. Recent testings show that Linux software execution times on the Red Pitaya development board do not exceed 75 ms, leaving plenty of time for other non-MCA related operations to be performed between subsequent reports. Speed is achieved by simply integrating the counts within the predefined mass channels that correspond to each of the cabin air components. This is in sharp contrast to the annual TGA mode of operation, which will perform pre-concentrations of cabin air samples on a daily basis for total of 10 minutes. Pre-concentration phase will then be followed by the MEMS injection of the concentrated sample into the QITMS sensor where it will be mass analyzed for VOCs and reported to the user. As long as the GC micro-column performs well, all of the VOCs found in Table will ideally arrive to the QITMS at separate time intervals; the detected signal will then resemble one of the mass spectra given in Fig. . Due to the degradation of the GC micro-column, analytes may start arriving to the QITMS in overlapping time intevals and their deconvolution must be performed by random walk procedure.

B. Maximum Entropy Random Walk

We begin our discussion with the test case in which all compounds are equipartially represented in an initial ion cloud of three different sizes, and we apply the random walk algorithm to recover these initial relative abundances for each compound (c) listed in Table . In Fig. , the initial relative abundances for each compound are represented with gray open symbols, such that the green dashed line marks the equipartial level. The shaded area denotes the 1-σ uncertainty interval assigned to the initial relative abundances. The relative abundances recovered by the random walk algorithm are given as red solid symbols, which tend to approach the open symbols as count statistics improve. It is evident that for sufficiently dissimilar compounds, the fully converged random walk algorithm may introduce an additional error into the recovered abundances, but this error falls within the error bars due to counting statistics. However, in the case of Xylene, the similarity in fragmentation of its isomers is too great to be completely factored out in the mass spectra that have unit mass resolution. In this case, recovered abundances can be as much as 0.25% off from the expected value of 4.545%. We note that the random walk method used here is multi-threaded, and that the final computation times can be shorter depending on how many cores are used. In single CPU runs the computational time required to recover relative abundances from an initial ion cloud containing 100 million fragments is about 3 min, and is mainly due to the manipulation of the large 8GB files.

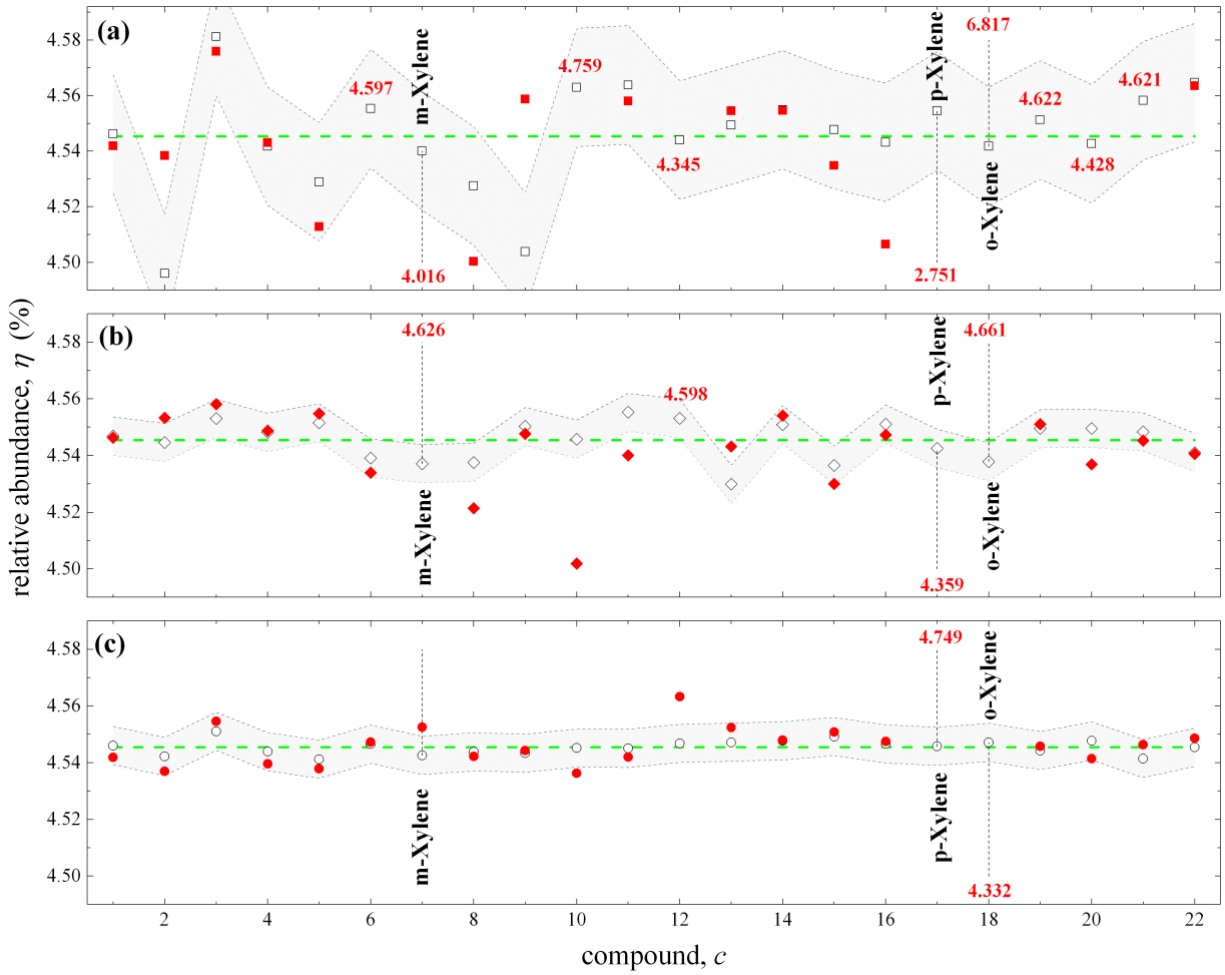

Figure 2 Error! No sequence specified.**: Effect of accumulated counting statistics.** *Reliability of random walk algorithm for unit mass resolution improves with the total number of detected ions: (a) 1 million, (b) 10 million, and (c) 100 million. Vertical dashed lines denote the recovered abundances for Xylene isomers.*

C. Minimum Entropy Random Walk

It is of particular interest to benchmark our computational method on the test cases where the initial ion cloud contains fragments of a single compound. We refer to these cases as minimum entropy with respect to the prior knowledge of the initial ion cloud composition. As the worst case scenario, we can then observe how the similarity in fragmentation patterns among the isomers of Xylene will cause the initially absent compounds to be erroneously reported at low levels in the recovered abundances. Presently, these results can be used to further quantify the extent to which particular counting statistics and the current computational algorithm may introduce the false positives. Breakdown of false positives for Xylene isomers is given in Table and for other compounds the level of false positives is at least an order of magnitude smaller. For each ion cloud, initially containing N_1=45454, N_2=10N_1 , or N_3=100N_1 fragments of a pure compound, we generate a unit mass resolution mass spectrum to which we then apply the random walk algorithm in order to extract the relative abundances of all compounds of interest.

Table 2 Error! No sequence specified.**: Effect of fragmental inter-dependence on propagation of false positives (in %)**

(c)	pure m-Xylene (c=7)			pure p-Xylene (c=17)			pure o-Xylene (c=18)		
	N_1	N_2	N_3	N_1	N_2	N_3	N_1	N_2	N_3
1	0.0041	0.0000	0.0004	0.0041	0.0184	0.0000	0.0608	0.0147	0.0000
2	0.0122	0.0220	0.0001	0.0122	0.0012	0.0001	0.0121	0.0013	0.0001
3	0.0056	0.0006	0.0000	0.0056	0.0006	0.0003	0.0056	0.0006	0.0004
4	0.0041	0.0004	0.0000	0.0206	0.0033	0.0000	0.0246	0.0045	0.0020
5	0.0041	0.0004	0.0000	0.0041	0.0004	0.0007	0.0041	0.0065	0.0000
6	0.0130	0.0000	0.0170	0.0065	0.0000	0.0000	0.0065	0.0000	0.0000
7	**97.5618**	**98.7637**	**99.8890**	**2.5556**	0.0008	0.0001	0.0000	0.0008	0.0001
8	0.5321	0.0000	0.0693	0.0000	0.0393	0.0001	0.0112	0.0219	0.0161
9	0.0052	0.0067	0.0001	0.0052	0.0005	0.0000	0.0051	0.0005	0.0000
10	0.0038	0.0004	0.0000	0.0039	0.0004	0.0000	0.0038	0.0004	0.0000
11	0.0087	0.0009	0.0016	0.0088	0.0009	0.0001	0.0087	0.0009	0.0016
12	0.0041	0.0004	0.0000	0.0041	0.0004	0.0000	0.0041	0.0004	0.0000
13	0.0116	0.0012	0.0057	0.0116	0.0012	0.0001	0.0116	0.0012	0.0001
14	0.0125	0.0004	0.0000	0.0042	0.0004	0.0000	0.0208	0.0004	0.0032
15	0.0054	0.0005	0.0000	0.0054	0.0005	0.0001	0.0054	0.0005	0.0001
16	0.0042	0.0004	0.0000	0.0042	0.0004	0.0000	0.0042	0.0004	0.0000
17	0.0000	**1.0679**	0.0001	**97.2281**	**99.8212**	**99.9920**	0.0000	0.0007	**0.9217**
18	**1.7485**	0.0006	0.0001	0.0000	0.0006	0.0000	**99.7808**	**99.9413**	**99.0430**
19	0.0056	0.1154	0.0078	0.0905	0.1069	0.0056	0.0056	0.0006	0.0114
20	0.0423	0.0170	0.0089	0.0141	0.0014	0.0001	0.0141	0.0014	0.0001
21	0.0062	0.0006	0.0000	0.0062	0.0006	0.0000	0.0062	0.0006	0.0000
22	0.0049	0.0005	0.0000	0.0049	0.0005	0.0006	0.0048	0.0005	0.0004

The main objective of this protocol is to arrive at the conclusion that the recovered abundances point to the presence of only one chemical compound, but because of the coarse resolution in the reference mass spectrum we expect that a certain level of similarity will exist among different species. This degree of similarity should be reduced with larger numbers of initial ions and finer resolutions of the reference mass spectrum, see Fig. . As expected, departures from 100% abundance decrease with the increase in number of detected ions. For example, when only the p-Xylene was present in the test mixture, the random walk predicted a 2.6% false positive for m-Xylene when the counting statistics were at level N_1, but almost none at level N_2 due to the appearance of less abundant fragments. Furthermore, for pure m-Xylene we have false positives above 1% due to p- and o-Xylene, which depend negligibly on the counting statistics. This is an indication that improvement for this isomer is only possible if the reference mass spectrum $\acute{R}^{exp}$ has sub-unit mass resolution. Although human and animal data show that all Xylene isomers or Xylene mixtures cause similar acute and chronic health and hazard effects, it is important to correctly identify particular Xylene isomers because they differ in their potency.

III.Summary

We have presented a progress report on the development of the software to be used for the analysis of the MS data produced by the S.A.M. instrument. At this time we focus only on the reporting accuracy for the trace VOCs. Obtained results will be used to benchmark requirements on the counting statistics when employing the MEMS

PCGC technology. The main finding is that the binning of detected ion counts (at low counting rates) into unit mass resolution histograms is too coarse for the parts of the mass spectrum below 100 Da, where fragmentation dissimilarities between different isomeric VOCs are most prominent. Finer mass resolution in this region of mass spectra will significantly improve the accuracy and efficiency of the random walk algorithm in recovering true relative abundances of Xylene isomers. In the case of the lowest level of counting statistics for pure p-Xylene isomer, the false positive response is at most 2.6%. One area for future study involves determining at which sub-unit mass resolution isomeric false positives are completely suppressed.

Acknowledgments

This work has been carried out at the Jet Propulsion Laboratory, California Institute of Technology, under the contract with the National Aeronautic and Space Administration. The authors thank S. Schowalter for his careful reading of the manuscript. © 2016. California Institute of Technology. Government sponsorship acknowledged.

References

[1]**Error! No sequence specified.**S.M.Madzunkov et al., "Progress Report on the Spacecraft Atmosphere Monitor", *46thInternational Conference on Environmental Systems*, Viena, Austria; abstract 284 (2016).

[2]**Error! No sequence specified.**NIST Standard Reference Database 1A, "NIST/EPA/NIH Mass Spectral Library with Search Program", *Data Version: NIST 14, Software Version: 2.2g*

[3]**Error! No sequence specified.**D. Nikolić et al., "Computer Modeling of an Ion Trap Mass Analyzer, Part I: Low Pressure Regime", *Journal of the American Society for Mass Spectrometry* 26(12): 2115-2124 (2014).

[4]**Error! No sequence specified.**S. Madzunkov et al., "Formation of Formaldehyde and Carbon Dioxide on an Icy Grain Analog Using Fast Hydrogen Atoms", *The Astrophysical Journal* 697(1):801-806 (2009).

[5]**Error! No sequence specified.**Agency for Toxic Substances and Disease Registry (ATSDR), "Toxicological Profile for Xylenes (Update)", *Public Health Service, U.S. Department of Health and Human Services*, Altanta, GA (2007).